

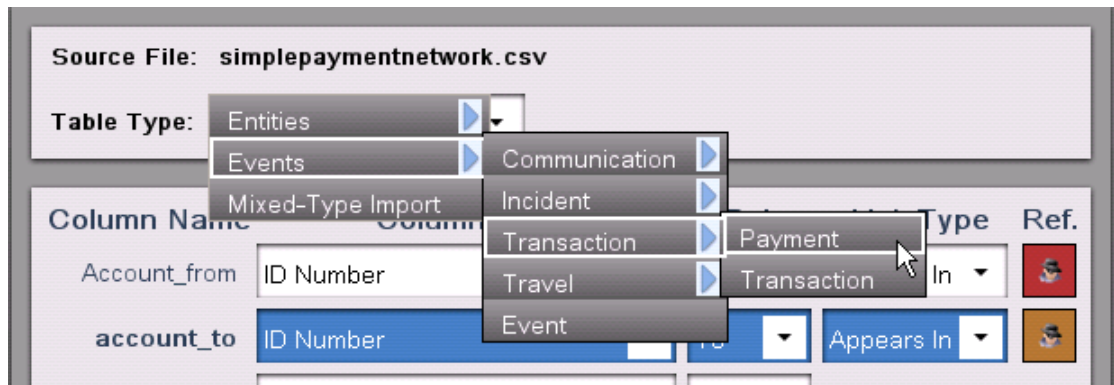
周报

1. 本周主要看了一些 palantir 的 online 的 PPT，着重在数据的底层模型的构建上。

要维护一个比较 general 的本体的构建方法，目前我自己的想法是：

- 1) 底层将 ontology 中的所有的内容分为: entity, event, document, media, property, link。其中 entity, event, document, media 都属于 object。每一个 object 都要维护一系列的 property (如果要使 property 足够灵活, property 的存储格式应类似于[property name, property value]的类似结构), 以及一系列的 related object。例如, 对于一个 entity, 需要维护[properties, related entities, related events, related documents, related media], 其他的也都类似。
- 2) 对于 property, 需要分为三种: simple property (just a simple value), composite property(for example, first name, last name.....), raw property(just an unparsed value). 每个 property 需要维护的有[parent object, child object, type, role, start time, end time.....]
- 3) 对于所有本体构建时信息, 都需要维护一个 data source record, 用于记录这个数据信息来自于哪个数据源的哪个位置。所以一个 data source record 需要保存该条数据在数据源中的第几行等信息。

Global 的 ontology 可以构建地很复杂, 但我认为我们现阶段不需要。现阶段我认为整个项目的流程是: 用户导入一个数据库或者 csv 之类的结构化数据源, 让用户选择每一列的类型 (event, entity.....等), 以及每一列的 column type, link type.....等。根据用户的这些输入构建出该数据源的 local ontology。每个 local ontology 构建好后, 再把 local ontology 合并到 global ontology 中。类似于:



这其中可能存在的问题是 local ontology 添加到 global ontology 时，应该需要名字之间的映射（如 local ontology 中的 name 在 global ontology 中可能是 enterprise name 之类的），不过这部分映射应该如何处理目前还没有考虑清楚。

之前看的好多资料都是采用 protégé 这个工具来构建本体，让王琦研究了工具，后面本体的构建应该要借鉴 protégé 的。

下周工作：

- 1.本周让大家都看了一些相关的内容，下周的前几天大家讨论一下，制定一下整个项目的大概的框架、流程、底层的数据结构。